zhejiang-yige

智能软件综合研究

实训报告



学期 2023-2024-1

班级 计算机科学与技术21(4)班

姓名 陈昊天

学号 2021329600006

教师 马明泽、吴婷婷

|  |  |  |
| --- | --- | --- |
| 学号： 2021329600006 |  | 姓名： 陈昊天 |

1. **实验目的和要求**
2. 掌握数据预处理相关操作；
3. 掌握深度学习模型构建、训练、评估、优化等过程；
4. **实验内容**

乳腺癌预测问题

Kaggle提供的威斯康星州乳腺癌数据集包含569个乳腺肿块样本，其中良性病例357个和恶性病例212个。每个样本包含1个ID号，10个乳腺肿块相关的重要特征，以及1个类别标签（良性、恶性）。通过数据预处理、数据集划分、模型构建、模型训练、模型评估等步骤，完成乳腺肿块是否癌变的预测任务。

1. **实验任务**

1.数据加载及预处理

数据加载及数据确认。要求展示头部数据、尾部数据、数据类型，确认数据完整性。

数据预处理。

（1）检查是否有重复数据并删除；

（2）检查是否存在缺失数据以及各特征缺失数据的比例，对缺失数据进行填充操作；

（3）将字符串和布尔字符转换成整数，完成编码操作；

（4）尝试分析特征之间的关系，将其可视化。

预处理数据保存。

2.数据集分割。将预处理后的数据集分割成训练集和测试集，其中训练集70%，测试集30%。

3.模型构建、训练、评估

分别利用决策树、支持向量机、随机森林三种模型对以上预处理后的乳腺癌数据进行训练、准确度评估、模型性能对比。

1. **实验过程和结果分析**

**1.数据加载及预处理**

# 加载数据

df = pd.read\_csv("data\_lib1.csv")

# 展示头部数据

print(df.head())

# 展示尾部数据

print(df.tail())

# 数据类型和完整性确认

print(df.info())

# （1）检查是否有重复数据并删除

df = df.drop\_duplicates()

# （2）检查是否存在缺失数据以及各特征缺失数据的比例

print(df.isnull().mean())

# 使用KNNImputer根据特征间的关系填充缺失数据

imputer = KNNImputer(n\_neighbors=5)

# （3）将字符串和布尔字符转换成整数

# 类别标签'diagnosis'是字符串需要被转化为整数，其余为数值特征

df["diagnosis"] = df["diagnosis"].map(

{"M": 1, "B": 0}

) # M:1 for malignant, B:0 for benign

df.iloc[:, 2:] = imputer.fit\_transform(df.iloc[:, 2:]) # 使用KNN填充缺失值，不包括ID和诊断结果

# （4）尝试分析特征之间的关系，将其可视化

sns.pairplot(df.iloc[:, 1:6], hue="diagnosis")

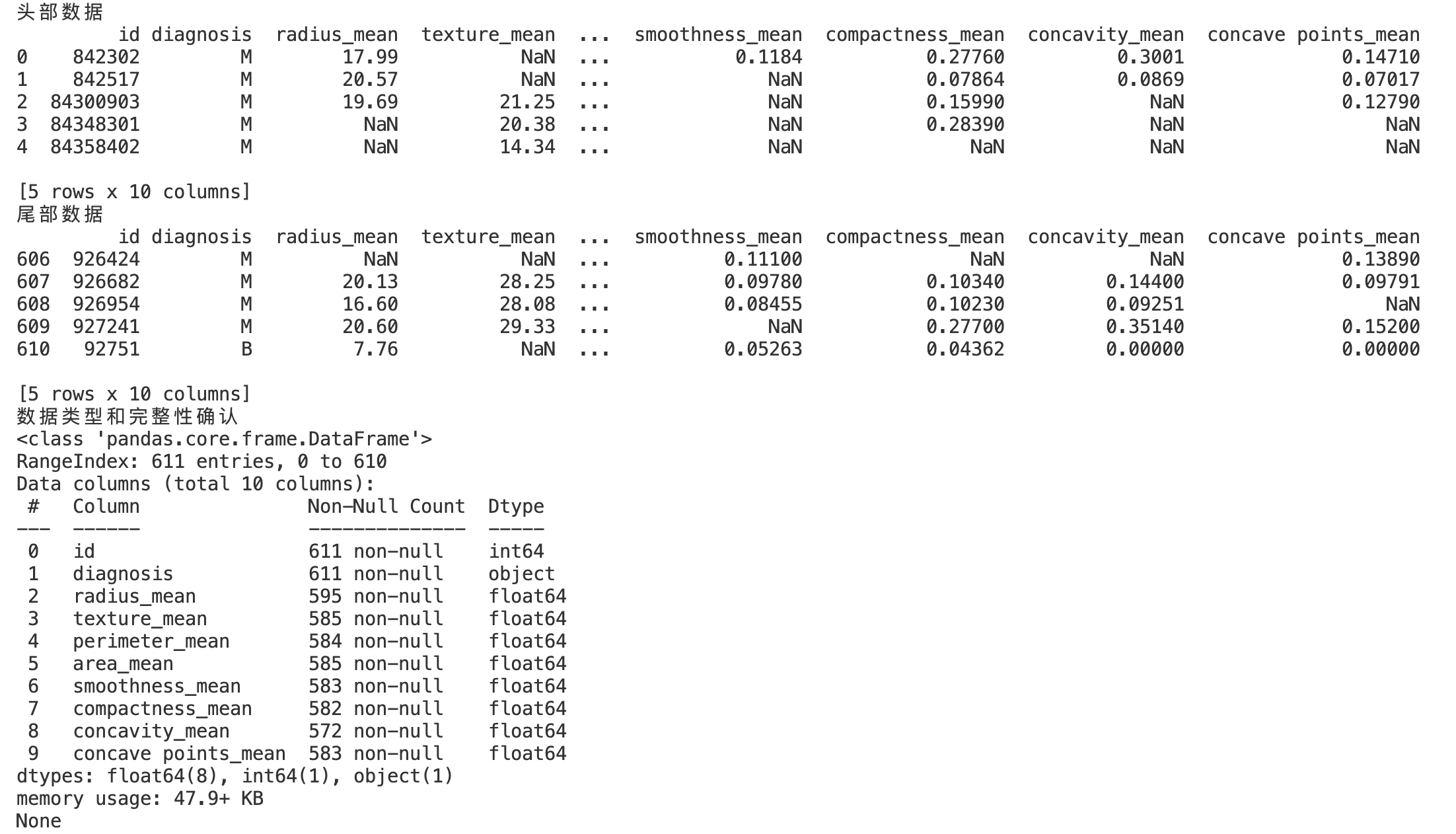
plt.show()

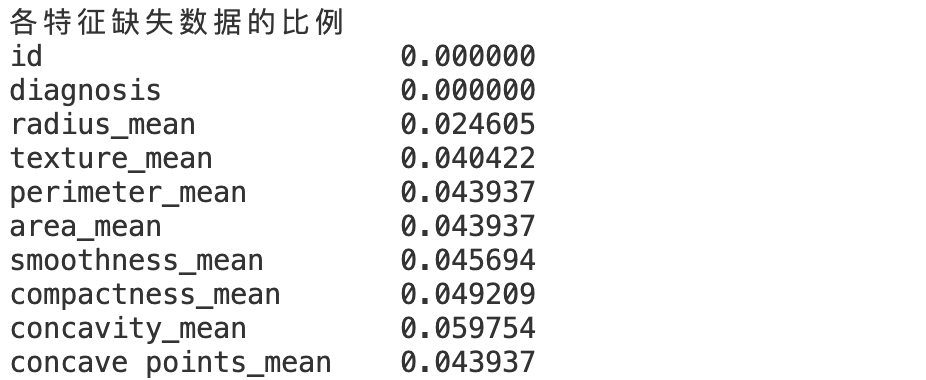
sns.pairplot(df.iloc[:, [1] + list(range(6, 10))], hue="diagnosis")

plt.show()

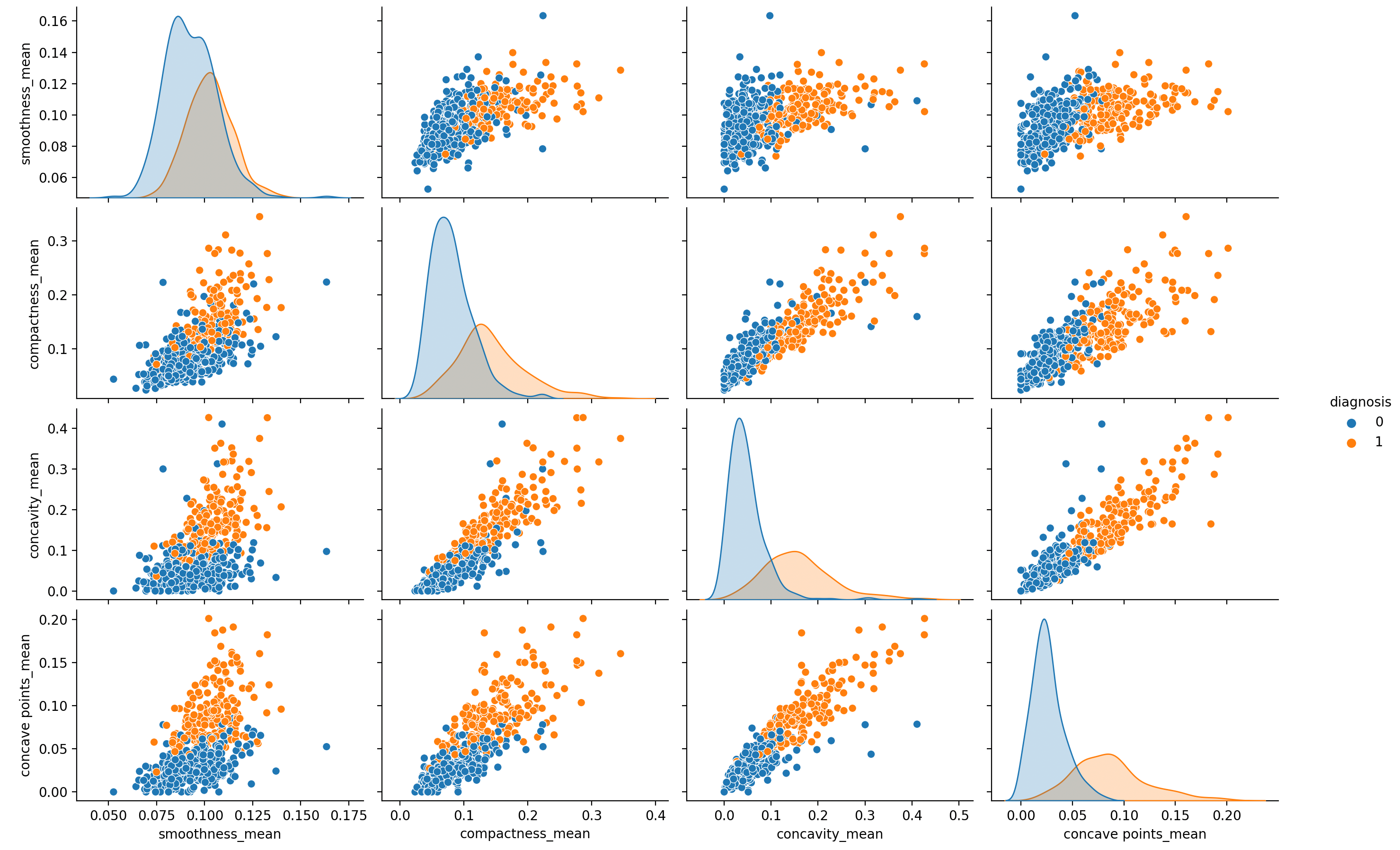
# 预处理数据保存

df.to\_csv("data\_lib1\_preprocessed.csv", index=False)









**2.数据集分割。**

# 分离特征和标签

X = df.iloc[:, 2:10] # 特征列

y = df["diagnosis"] # 标签列

# 分割数据集

X\_train, X\_test, y\_train, y\_test = train\_test\_split(

X, y, test\_size=0.3, random\_state=42

)

**3.模型构建、训练、评估**

# 决策树模型

dt\_model = DecisionTreeClassifier(random\_state=42)

dt\_model.fit(X\_train, y\_train)

y\_pred\_dt = dt\_model.predict(X\_test)

accuracy\_dt = accuracy\_score(y\_test, y\_pred\_dt)

print("Decision Tree Accuracy: {:.2f}%".format(accuracy\_dt \* 100))

print(classification\_report(y\_test, y\_pred\_dt))

# 支持向量机模型

svm\_model = SVC(random\_state=42)

svm\_model.fit(X\_train, y\_train)

y\_pred\_svm = svm\_model.predict(X\_test)

accuracy\_svm = accuracy\_score(y\_test, y\_pred\_svm)

print("SVM Accuracy: {:.2f}%".format(accuracy\_svm \* 100))

print(classification\_report(y\_test, y\_pred\_svm))

# 随机森林模型

rf\_model = RandomForestClassifier(random\_state=42)

rf\_model.fit(X\_train, y\_train)

y\_pred\_rf = rf\_model.predict(X\_test)

accuracy\_rf = accuracy\_score(y\_test, y\_pred\_rf)

print("Random Forest Accuracy: {:.2f}%".format(accuracy\_rf \* 100))

print(classification\_report(y\_test, y\_pred\_rf))

# 可视化预测结果

def plot\_confusion\_matrix(y\_true, y\_pred, title):

cm = confusion\_matrix(y\_true, y\_pred)

sns.heatmap(cm, annot=True, fmt="d", cmap="Blues")

plt.title(title)

plt.xlabel("Predicted")

plt.ylabel("Actual")

plt.show()

plot\_confusion\_matrix(y\_test, y\_pred\_dt, "Confusion Matrix - Decision Tree")

plot\_confusion\_matrix(y\_test, y\_pred\_svm, "Confusion Matrix - SVM")

plot\_confusion\_matrix(y\_test, y\_pred\_rf, "Confusion Matrix - Random Forest")

# 模型性能对比

models = ["Decision Tree", "SVM", "Random Forest"]

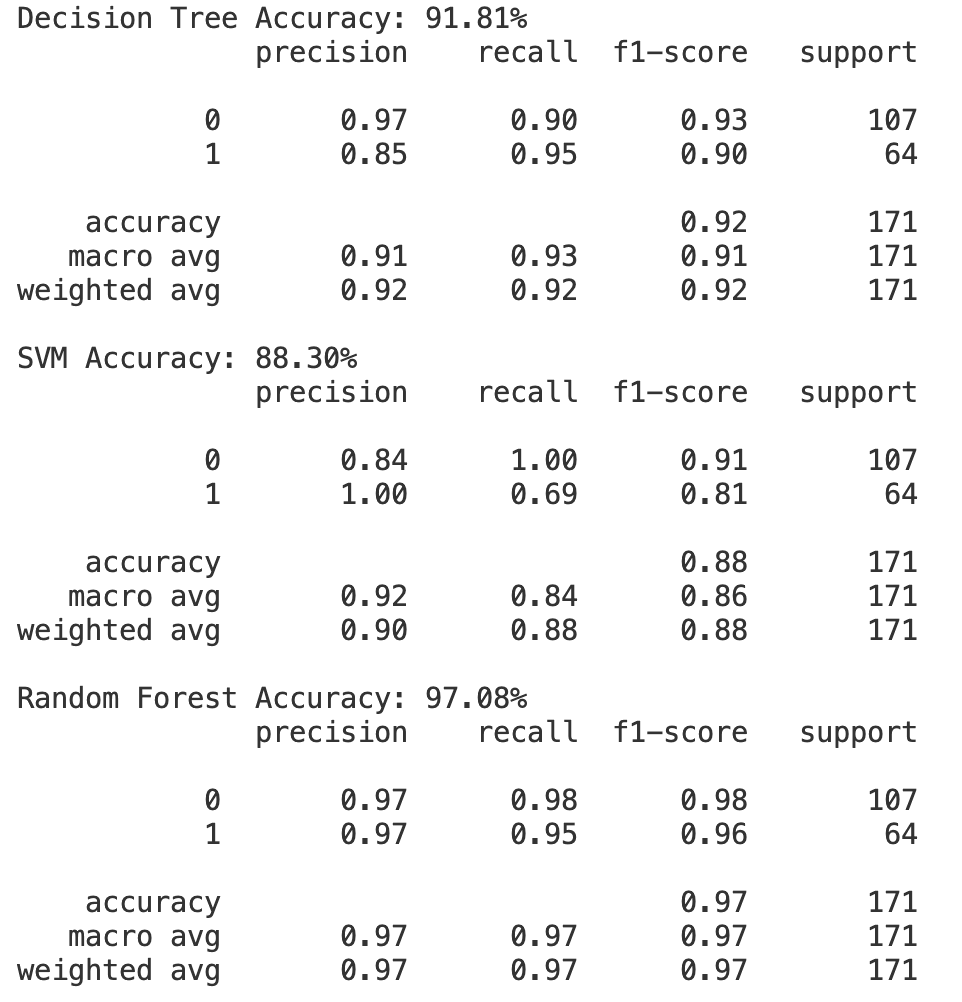
accuracies = [accuracy\_dt, accuracy\_svm, accuracy\_rf]

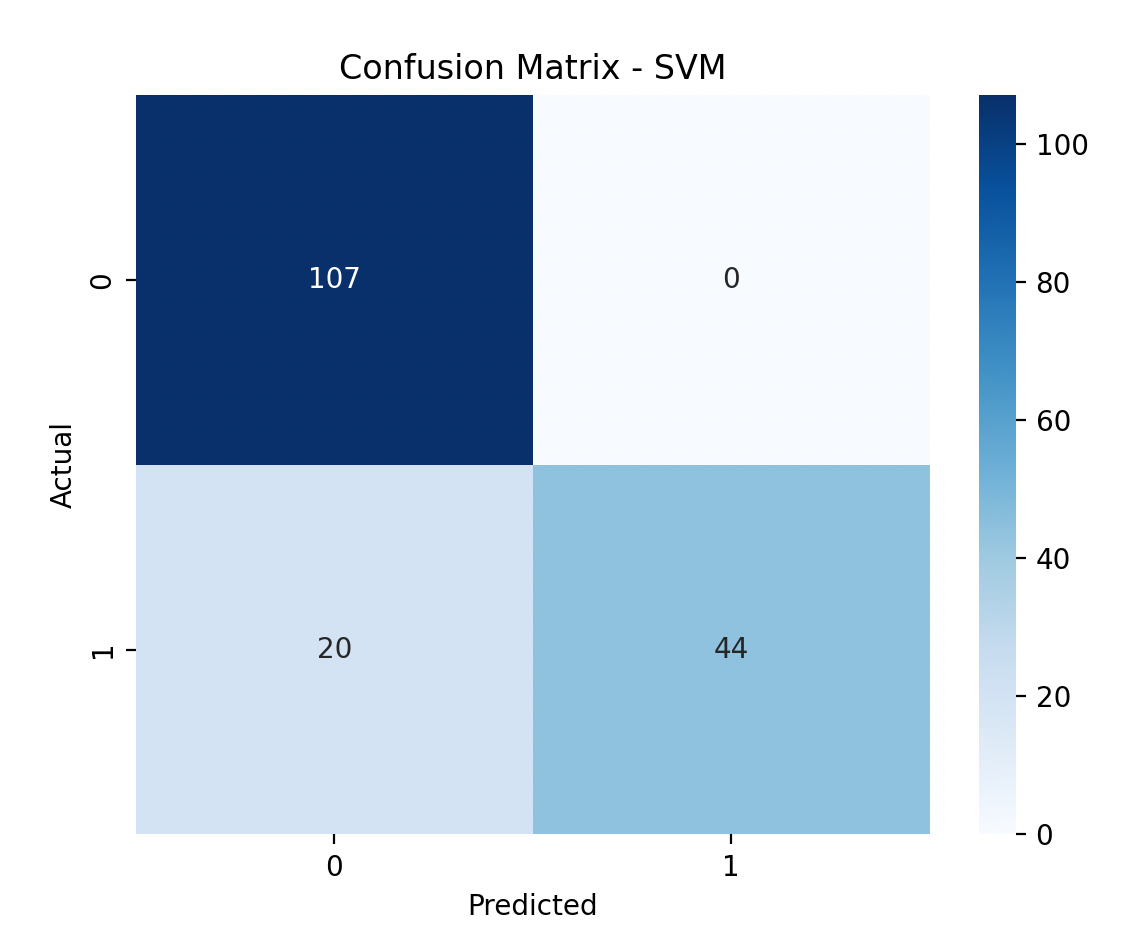
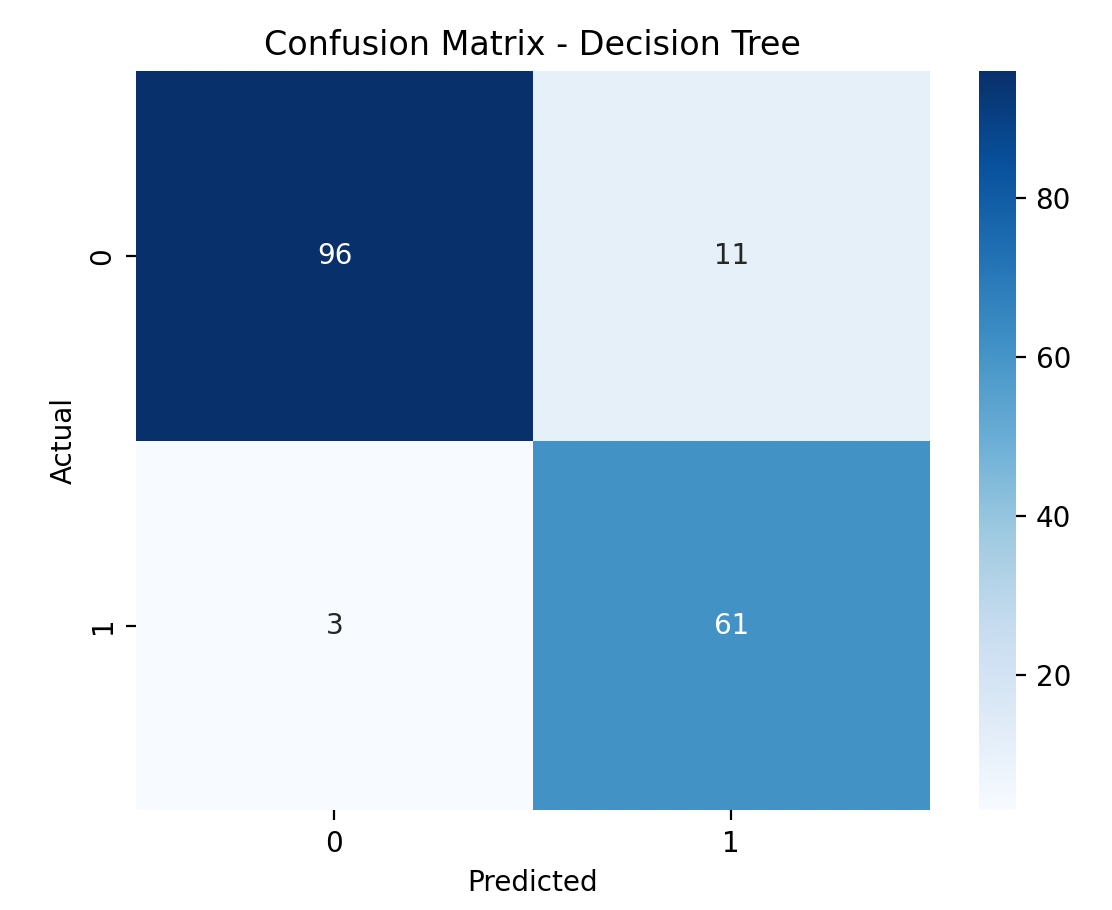
plt.bar(models, accuracies)

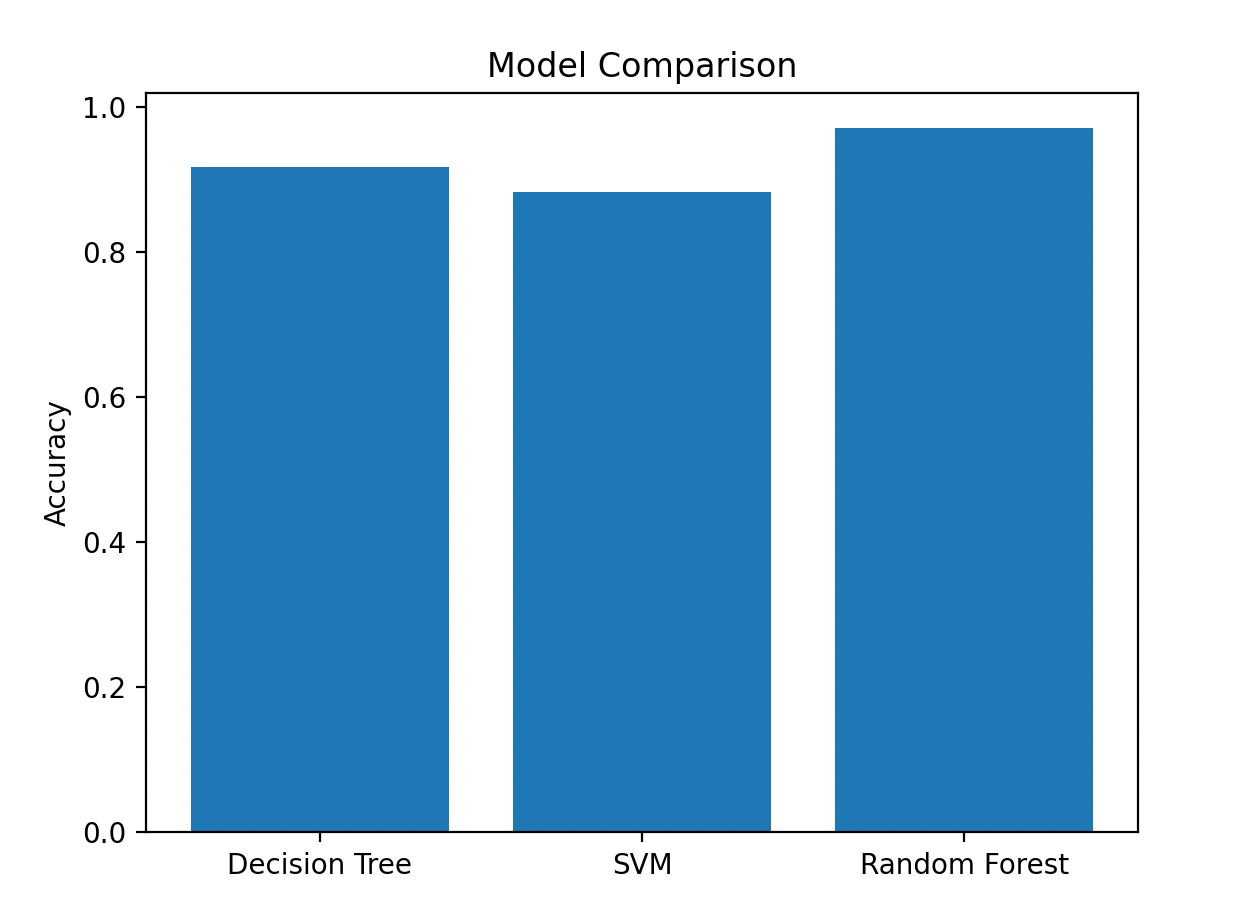
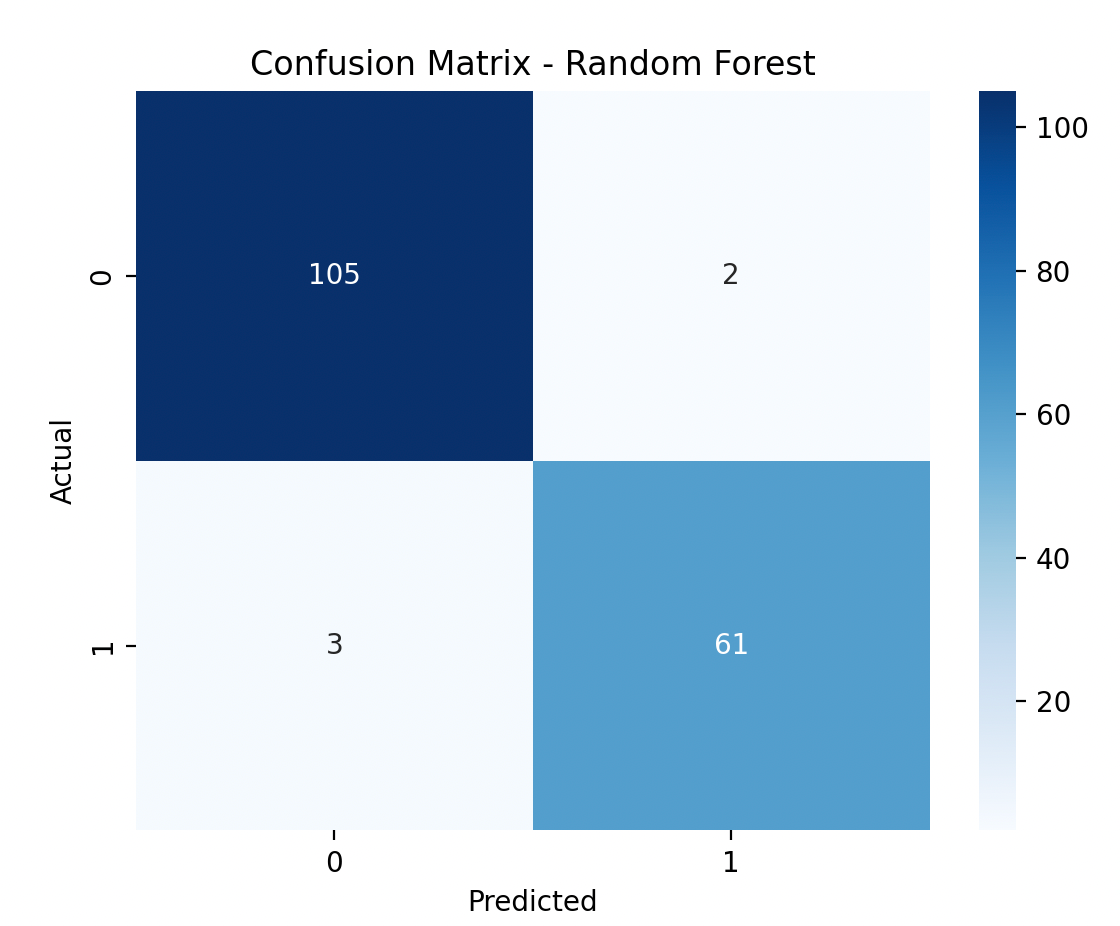
plt.ylabel("Accuracy")

plt.title("Model Comparison")

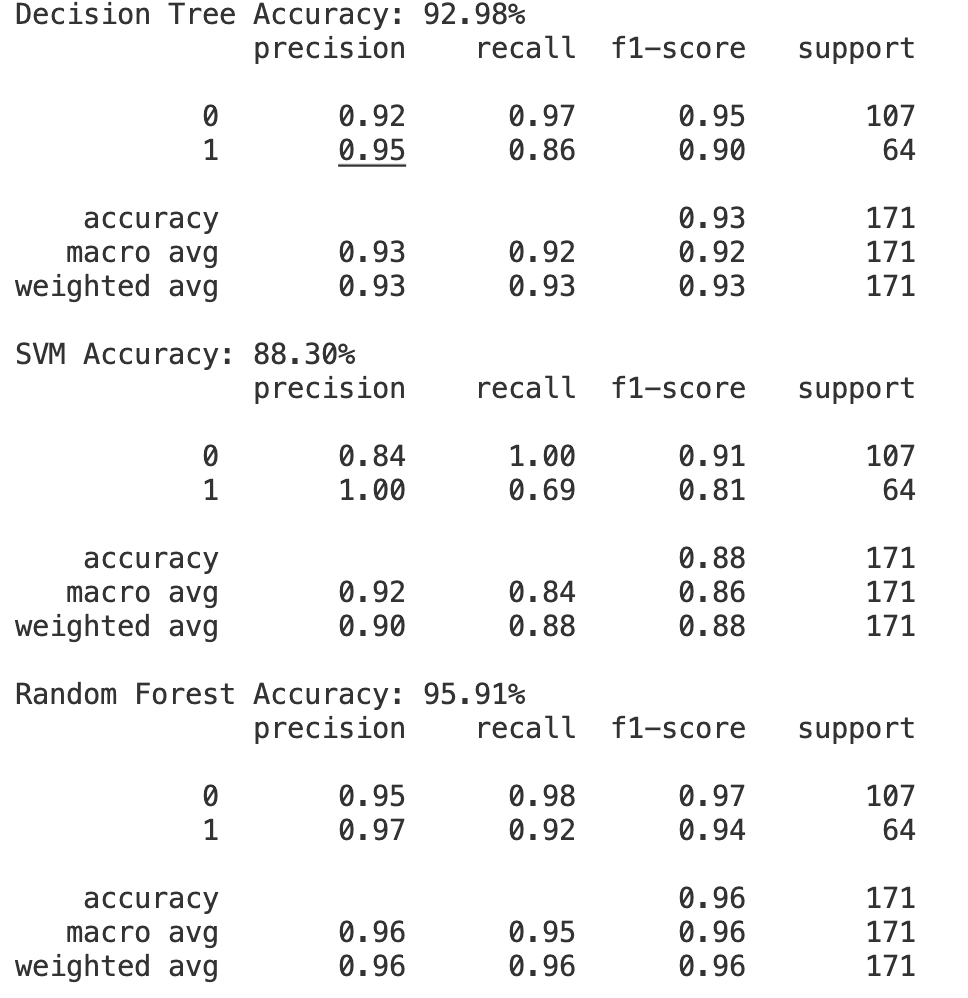
plt.show()







实验还研究了concave points\_mean特征对模型精确度的影响。在数据集去除该项数据后，模型的准确度发生了一定变化。



1. **实验总结**
   1. **总结**

本次实验的主要目的是通过实际操作学习和掌握数据预处理、深度学习模型的构建、训练、评估及优化过程，并应用于乳腺癌预测问题。实验使用了Kaggle提供的威斯康星州乳腺癌数据集，并通过决策树、支持向量机和随机森林三种不同的机器学习模型进行了训练和评估。

在数据预处理阶段加载了数据并对其进行了初步的检查，包括展示数据的头部和尾部、确认数据类型和完整性。接下来进行了去重、填充缺失值、数据编码的操作，并尝试分析了特征之间的关系并将其可视化。预处理后的数据被保存，以便后续使用。

数据集分割步骤中，将预处理后的数据集分为训练集和测试集，比例为70%和30%。这一步骤是为了确保模型能在未见过的数据上进行有效的预测。

在模型构建、训练和评估阶段，分别对三种模型进行了训练并评估了它们的准确度。可视化了预测结果并对不同模型的性能进行了对比。此外，实验还研究了concave points\_mean特征对模型精确度的影响，发现去除该特征后模型的准确度发生了变化，说明该特征对模型预测性能有一定的影响。

根据实验结果，可以得出结论，数据预处理是模型构建的重要步骤，不同的模型对于特定问题的适应性有差异，而特定的特征对模型的准确性也有显著影响。通过实验，我们加深了对机器学习模型构建和评估流程的理解，并获得了实际操作经验。

* 1. **心得体会**

通过本次实验，我深刻体会到了数据预处理在机器学习中的重要性。高质量的数据是模型良好性能的基础，缺失值填充、数据编码等预处理操作对于后续模型的训练有着直接的影响。

实验中，我了解并实践了不同机器学习算法的应用，包括决策树、支持向量机和随机森林。通过比较这些模型的性能，我认识到了选择合适模型的重要性，以及针对特定问题调整模型参数的必要性。

特征选择对模型性能的影响也给我留下了深刻的印象。实验中，concave points\_mean特征的移除对模型准确度产生了影响，这提醒我在构建模型时，需要仔细考虑哪些特征是真正对预测结果产生影响的。

本次实验不仅让我掌握了理论知识，更重要的是让我通过实践学会了如何应用这些知识解决实际问题，这将对我的未来学习和工作产生积极的帮助。